ANALYSIS

4(15), July - September, 2018

ISSN 2394–8558 EISSN 2394–8566

Climate Change

Exploring forest aboveground biomass estimation using landsat, forest inventory and analysis data base

Seyed Omid Reza Shobairi[1☼], Vladimir Andreevich Usoltsev[1,2], Viktor Petrovich Chasovskikh[2], Mingyang LI[3]

This paper, aims to calculate the various remote sensing based models namely multiple linear regression (MLR), k-Nearest Neighbor (k-NN), bagging (Bagging) and random forest (RF); that were established using 9 vegetation index and three terrain variables. Five indicators of correlation coefficient (COR), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (%RAE), root relative squared error (%RMSE) were figured out to evaluate the performance of the four models using ten-fold cross validation method. Then the model with the best performance was applied to predict the dynamics of forest aboveground biomass during 1993 to 2013. Results show that among the four models, the prediction accuracy of random forest is the highest, followed by k-NN method, while the accuracy of MLP is the lowest. The terrain factors including elevation and slope, soil conditions (e.g. brightness, wetness), vegetation growth conditions (e.g. vertical vegetation index, effective leaf area index) are the enforcing factors impacting regional forest aboveground biomass. During 1993 to 2013, the unit forest biomass in study area decreased from 34.68 Mg/ha in 1993 to 32.59 Mg/ha in 2003, then increased to 44.65 Mg/ha in 2013. The spatial distribution pattern of forest aboveground biomass had experienced a change from high aggregation to fragmentation from 1993 to 2013. The spatial hot/cold spots analysis predicted that the cold spots with most gentle changes of forest aboveground biomass during research period were mainly distributed in the northern rocky mountains with inconvenient traffic conditions, high forest cover and less human disturbance, while the hot spots with most dramatic changes located in the southern Guan River valley with good traffic conditions, high population density and gentle slope.

INTRODUCTION

Since forest biomass stores large amounts of carbon, the amounts of biomass in forest can reflect the functions of forest ecosystem and are closely related to the production level of forest [1]. The estimation of forest biomass is the basis for the research of ecology material recycling [2]. The analysis of the driven factors of the change of forest biomass can provide scientific guidance for the making of sustainable forest management measures. In collectively owned forest region of China, fast growing forest plantations with high production and short harvest rotation account for higher proportion of forest area. Since 1981, forest in the collectively owned forest region has been divided into small plots and allocated to the management unit of household. Therefore, forest management levels have been greatly impacted by forest ecological benefit compensation, cutting quota, and other policies and regulations. Since then, forest in collectively owned region has been experiencing rapid change both in spatial distribution pattern and forest structure [3]. Estimation of forest biomass and analysis of driven factors at regional scale are the basis for the precise calculation of carbon sink [4], and can provide scientific guidance for the making of sustainable forest management measures.

[1]Ural State Forest Engineering University, Sibirskii trakt 37, Yekaterinburg, 620100 Russian Federation; [2]Botanical Garden, Russian Academy of Sciences, Ural Branch 8 Marta str., 202a, Yekaterinburg, 620144 Russian Federation; [3]College of Forestry, Nanjing Forestry University, Nanjing 210037, China.
☼*Corresponding Author:* Ural State Forest Engineering University, Sibirskii trakt 37, Yekaterinburg, 620100 Russian Federation; E-mail: Omidshobeyri214@gmail.com, Telephone:+8(343) 254-61-59

Monitoring of forest biomass is conducted through quantitative analysis of the change of forest biomass during a specific period of time. Commonly used methods include forest continuous inventory, flux observation, simulation modeling method, remote sensing based estimation, etc [5]. Remote sensing based methods estimate the amount of biomass using empirical models derived from statistics analysis of remote sensing data and field survey forest biomass. There are several ways to monitoring dynamics of forest biomass using remote sensing based methods, such as regression, nonparametric imputation, non-parametric regression tree method.

Among multiple remote sensing based methods, regression method is widely used in the estimation of forest biomass. Labrecque et al. [6] extracted vegetation index from Landsat TM raw spectral band data as independent variables to fit the polynomial and multivariate linear regression models in order to predict the forest aboveground biomass in Newfoundland, Canada. The basic idea of nonparametric imputation method, like k-nearest neighbor algorithm (k-NN), gradient nearest neighbor algorithm (GNN), is to assign an image pixel with the average value of its nearest neighbors. Due to the huge analytical and operational flexibility, nonparametric regression trees have been gradually paid more attention in the prediction of forest structure parameters. Blackard et al. [7] mapped the spatial distribution of forest aboveground biomass in Continental USA using regression tree method with remote sensing data of MODIS and other ancillary materials. Moisen and Frescino [8] compared the performance of regression tree method with several other statistical models when they estimated the forest biomass of 5 ecological zones in Western America. Recent researches show that random forest has very good adaptability in predicting forest parameter such as succession stage, species distribution, degree of canopy damage caused by the forest fires, etc.

Compared with the research work done by international experts, Chinese remote sensing based estimation of forest biomass is limited to single model prediction within a short temporal period. Xing et al. [10] established a regression model to predict larch forest aboveground biomass using Landsat ETM+ data. Guo et al. [11] established a multivariate regression model to predict forest biomass in southern slope of Greater Xing'an Mountains using forest continuous survey plot data and vegetation index derived from Landsat TM. Chen et al. [12] evaluated the adaptively of a k-NN prediction model forest aboveground biomass by ten-fold cross evaluation using Landsat TM data and forest continuous plot data in small area of Jilin Province, China.

Study Area

Xixia County (111º01′-111º46′E, 33º05′-33º48′N), is located in the southwest part of Henan Province, China (Fig.1), with a total population of 450,000 and land area of 3,454 square kilometers. The mountainous area occupies 80% of the total land area of the county, while the crop land, water and built up combined together occupy remaining 20%. It is high and medium altitude mountains in the north, low elevation hills in the south. The county is situated in the north subtropical monsoon zone, with a temperate climate, moderate rainfall, and sufficient sunshine. The annual mean temperature is 15.2℃, the annual average precipitation is 830 mm, and the average sunshine hours are 2019 h per year. There are 526 rivers and streams in the county, such as Guanhe River, Qihe River, Xiahe River, Shuanglong River and Danshui River. The Guanhe River, which is the longest river and belongs to the Danjiang River System of the Yangtze River basin, runs across the whole county from north to south.

Xixia County boasts abundant natural resources and regional forest coverage is 76.8%. It is an important forest county in Henan Province with largest forest area and stock volume. There are 3.96 million mu of forest (1 mu = 0.0667 hectares) and major tree species are *Pinus massoniana*, *Cunninghamia Lanceolata*, *Quercus variabilis* and *Liquidambar formosana*. It also has 1,328 species of medicinal herbs, 49 types of verified mineral reserves, and average annual water resources of 1.29 billion cubic meters. The county also contains four national and provincial-level nature reserves, which account for 22.2% of the county's total land area.

Located in the transition belt between north subtropical and temperate zone, Xixia County has good growing conditions for both north and south China tree species. Zonal natural vegetation type is evergreen broad-leaved forest and deciduous forest with more than 75 families and 450 species of trees and shrubs. Among them, the dominant evergreen tree species are pine (*Pinus massoniana*), fir (*Cunninghamia Lanceolata*), Armand Pine (*Pinus armandi*), Chinese pine (*Pinus tabulaeformis)*, etc. The major deciduous broad-leaved species are Quercus (*Quercus variabilis*), sweetgum (*Liquidambar formosana*), catalpa trees (*Catalpa bungei*), etc. There are 128 kinds of forest by product and 1,380 kinds of Chinese herbal medicines. Among multiple cash forest species, kiwi (*Actinidia chinensis*), dogwood (*Sieb etzucc*), tung (*Vernicia fordii)*, lacquer are known as "Four Treasures" of Xixia.

RESULTS AND ANALYSIS

Models evaluation

The plots with the land type of crop land, water, built-up were deleted from the dataset of 217 fixed plots. In this paper, remote sensing estimation models were established with the dependent variables of volume converted biomass and independent variables of 12 eco-environmental factors WEKA 3.7.12 was used to establish the models and calculate the validation indexes. To reduce the correlation between the independent variables, the method of M5 was chosen to select variables in the model of the MLP. When using k-NN method, we selected neighborhood k=3. For RF method, the number of feature variables was 3 and the number of generated trees was 500. Ten-fold cross validation method was used to verify the accuracy of the 4 remote sensing estimation models. In the modeling processes during the 5 periods from 1993 to 2013, the order of prediction accuracy for the four models remained unchanged: random forest was the highest, followed by k-NN method and bagging method, while the accuracy of MLP was the lowest Therefore, the performance of the four models were evaluated by using the average values of 5 validation indexes (Table 2).

Seen from Table 2, among the four models, the prediction accuracy of RF was the highest, followed by k-NN method and bagging method, while the accuracy of MLP was the lowest. The study area is situated in the transition belt between China's subtropical zone and warm-humid zone. In Xixia, the variation in elevation is dramatic, the rainfall is rich and the vertical distribution of forest is very obvious. As the relationship between forest biomass and many environmental factors, such as elevation, slope, wetness, greenness, is not linear, the overall performance of MPL was the lowest. In k-NN, the average value of the attributes of these neighbors is assigned to the sample by finding a sample of k nearest neighbors, thus the properties of the sample are obtained. As a lazy algorithm, k-NN can be used to maintain the covariance structure of the explanatory variables, so the prediction accuracy of this algorithm is higher than that of MLP. Bagging algorithm takes the idea of bootstrap method, and uses the simple sampling approach to take samples from population, forming a number of training samples and multiple tree models. The results are predicted on the basis of majority voting rule. Compared with the bagging algorithm, random forest is not only used for sampling samples, but also for sampling variables, so the performance of RF is best. Bagging and random forests will generate multiple tree models, and then the method of combination forecasting is applied, the forecasting accuracy is higher than classification and regression tree (CART) which only uses a single learning. Therefore, the prediction accuracies of these three integrated learning methods (RF, k-NN, Bagging) with many learners are higher than MLP.

Enforcing variables

Since the prediction accuracy of RF is the highest, this model was applied to select enforcing variables of forest biomass. Two indicators of relative importance of mean decrease accuracy (% IncMSE) and node purity (IncNodePurity) were calculated to evaluate the importance of 12 independent variables (Figure. 2) Mean decrease accuracy is used to measure the decreased degree of the random forest prediction accuracy. If a variable has a big % IncMSE value, the variable is more important than others. Node purity is measured by Gini Index which is the the difference between RSS before and after the split on that variable. Likewise, a bigger value of IncNodePurity indicates the variable is more

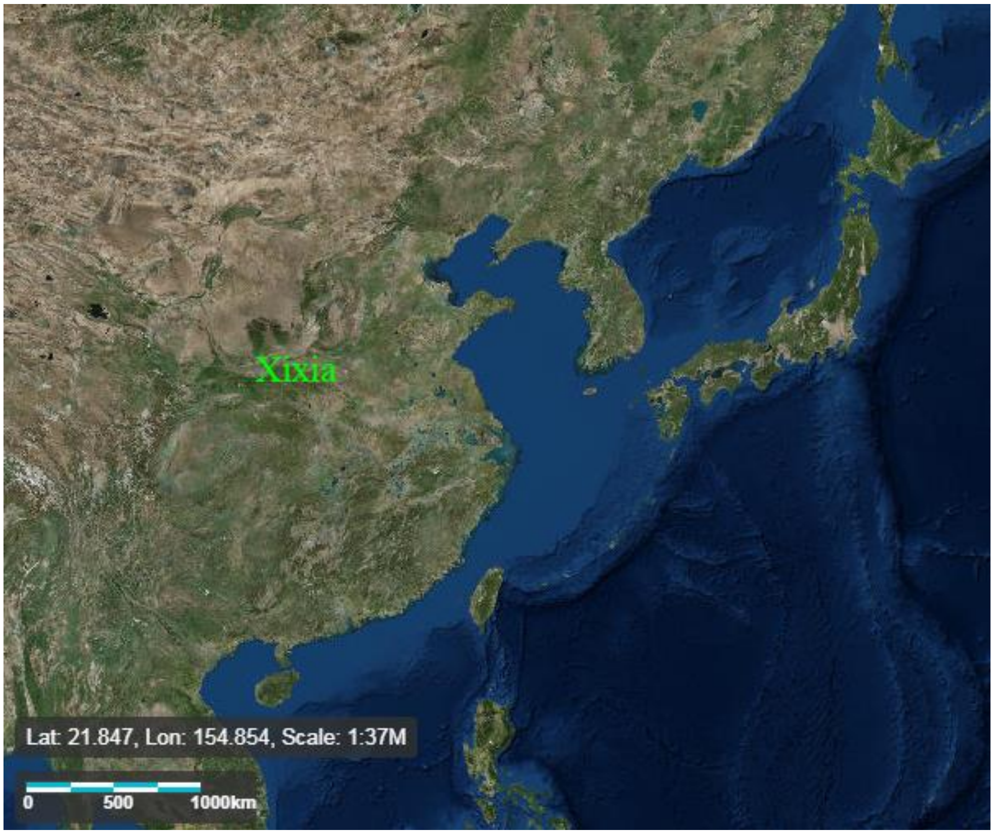

Figure 1 Location of study area (Xixia), Henan, China

Table 1 Parameters for forest volume and biomass exchange

No	Species/Species group	a	b
1	*Platycladus orientalis*	0.6129	46.1451
2	*Pinus massoniana*	0.5101	1.0451
3	*Cunninghamia lanceolata*	0.5371	11.9858
4	Quercus spp.	1.1453	8.5473
5	Broad-leaved hardwood	0.7564	8.310
6	*Populus* spp.	0.9810	0.0040
7	*Pinus tabulaeformis*	0.7554	5.0928
8	Mixed broad-leaved forests	0.8392	9.4157
9	Mixed coniferous forests	0.5894	24.5151
10	Mixed coniferous and broad-leaved forests	0.7143	16.9152
11	Paulownia spp.	0.8956	0.0048
12	Broad-leaved softwood 301603	0.4754	30.6034

important than others.

Ranked by the relative importance of mean decrease accuracy of 12 environmental variables, elevation, and effective leaf area index (SLAVI), brightness index, wetness index and vertical vegetation index (PVI) are enforcing variables affecting the forest aboveground biomass in the study area. Likewise, ranked by the values of the node's purity, elevation, effective leaf area index (SLAVI), slope, brightness index and greenness index are the important factors. Considering these two indicators, the terrain factors including elevation and slope, soil conditions (e.g. brightness, wetness), the vegetation index (vertical vegetation index, effective leaf area index), altogether 6 factors are the six enforcing variables impacting regional forest biomass. In an area with high altitude and steep slope, elevation, slope and other terrain factors affect the distribution of tree species and soil nutrient, which have direct impact on forest aboveground biomass. The brightness and wetness are closely related to the soil illumination environment and hydrological conditions and are the main factors that reflect the forest stand conditions. Therefore, they are important factors affecting the forest aboveground biomass. The vertical vegetation index (PVI) and effective leaf area index (SLAVI) are the key parameters which can reflect the vegetation growth status and determine the material and energy exchange between the forest ecosystem and the atmosphere, thus becoming the 2 important environmental variables affecting forest biomass.

Dynamics of forest aboveground biomass

The results show that there is a strong correlation among 5 validation indexes. The higher the correlation coefficient (COR) is, the lower the mean absolute error (MAE), the relative error (RAE), the root relative squared error (RRSE) and the root mean squared error (RMSE).

Therefore, the correlation coefficient (COR) of the predicted values and the measured values is used to evaluate the prediction accuracy of the four models. According to the principles of statistic, if COR lies between 0.9 and 1.0, two variables is extremely correlated; if COR lies between 0.7 and 0.9, highly correlated; if COR lies between 0.5 and 0.7, significantly correlated; if COR lies between 0.3 and 0.5, moderately correlated; if COR is less than 0.3, lowly correlated. Among the 4 models, only the average COR of the random forest model during five periods reached the level of high correlation. As a result, the model of random forest was applied to estimate the regional forest aboveground biomass from 1993-2013 (Fig. 3). Calculation results showed that the forest aboveground biomass in 1993, 1998, 2003, 2008 and 2013 is 34.68 Mg/ha, 33.66 Mg/ha, 32.59 Mg/ha, 36.89 Mg/ha, 44.65 Mg/ha, respectively. For convenience, the forest biomass in study area was divided into three categories: low (<20 Mg/ha), medium (20-50 Mg/ha), high (>50 Mg/ha).

Calculation results showed that during 1993 to 2013, the unit BIOMASS in study area decreased first (1993-2003) and then increased (2003-2013). Seen from the Fig. 3, the forest with high forest biomass (>50Mg/ha) presented the same change trend with the average unit forest biomass, while the area percent of forest with low forest biomass (<20Mg/ha) changed in the opposite direction. The area percent of forest with medium forest biomass (20-50Mg/ha) showed a trend of steady increasing from 1993 to 2013. During the period from1993 to 2013, there was a close relationship between the dynamics of regional forest biomass and the change of social conditions, forestry regulations and national macro-economic policies. Before 2003, the overall economic development level in Xixia was low. Deforestation and conversion of forest into crop land were the main channels of farmer’s income and food. This large scale forest destruction phenomenon was widespread in the stony mountainous area with high forest coverage and steep slope where economy was backward, resulting in a slowly decrease trend of forest biomass in the study area from 1993-2003. When the unit forest biomass had been decreasing from 1993 to 2003, the area percent of forest with low forest biomass had steadily increased, while the forest with high forest biomass decreased. From 2002, the project of Returning Farmland to Forest has been successfully implemented in the county. From 2004, in line with China’s South-North Water Diversion Project, forest cutting has been strictly forbidden in natural forest used for water conservation in Xixia. As a result, the phenomenon of deforestation and land reclamation in the northern and central mountainous area was dramatically declined. In 2008, Xixia County issued a document named “Implementation Plan of Ownership Reform of Collective Forest", which greatly improve the motivation of the majority of farmers to plant cash forest in the hilly land. Under the influence of urbanization, Xixia County has carried out large scale construction project of ecological corridor and village green space in the middle Guan River valley and the southern hilly basin with good traffic conditions, high population density since 2008. Since then, degraded forest and open woodland located in the hilly land with low elevation and high population density have been largely transformed into cash forest and recreational forest with higher forest biomass. Therefore, from 2003 to 2013, the area percent of forest with low forest biomass showed a significant decreasing trend, while the forest with high forest biomass steadily increased year by year. It should be noticed that, as the biggest forest stock volume county in Henan Province, the area percent of forest with medium forest biomass in the county showed a steadily increasing trend from 1993 to 2013.

Change of spatial pattern of forest biomass

Spatial autocorrelation analysis and spatial hot/cold spots detection were used to analyze the change of spatial pattern of forest biomass from 1993 to 2013 (Fig. 4). The spatial autocorrelation of biomass in the study area was measured using Moran's *I*. Moran's *I* is a weighted correlation coefficient used to detect departures from spatial randomness. Moran's *I* is used to determine whether neighboring areas are more similar than would be expected under the null hypothesis. Negative values indicate negative spatial autocorrelation and the inverse for positive values. Value of Moran's *I* ranges from −1 (indicating perfect dispersion) to +1 (perfect correlation)[22]. A zero value indicates a random spatial pattern. The spatial hot/cold spots detection attempts to find the sub-regions with attribute values significantly different from other regions in the study area, which are considered to be the abnormal regions, such as the regions with very low or extremely high forest biomass [23].

It can be seen from Fig. 4, the spatial aggregation of forest biomass in Xixia County had been decreasing continuously from 1993 to 2013. In the first decade of 1993 to 2003, the spatial aggregation of forest biomass decreased slowly, and then decreased dramatically in the second decade.

During the period of 1993 to 2003, deforestation and land reclamation in under developed northern and central mountainous area were widespread, causing forest biomass in these areas decreasing. A the same time, motivated by higher economic rate of return, farmers increased the planting area of cash forest in valley basin and hilly land, making the gap of forest biomass in different areas be narrowed and the spatial aggregation of forest biomass be decreased. With the fully implementation of polices of returning farmland to forest and logging ban on natural forest, the decreasing trend of forest biomass in northern and central mountainous area in Xixia County has been checked since 2002. With the advancement of the ownership reform of collective forest and the large-scale construction project of ecological corridor and village green space, the forest biomass and forest coverage in the middle of Guan River valley and the southern hilly basin with good traffic conditions and high population density, have been greatly increased, causing the spatial aggregation of forest aboveground biomass at county level decreased significantly from 2008 to 2013.

Through cold/hot spots analysis, the spatial aggregation of the standard deviation (SD) of forest biomass form 1993-2013 in study area was divided into four kinds: high value points (hot spots, HH), low value points (cold spots, LL), points with high value surrounded by low value points (HL), points with low value surrounded by high value points (LH) (Figure 5).

The average elevation, average slope of the hot spots and cold spots, and the average value of night light of DMSP/OLS imagery were extracted to analyze the differences of terrain factors and economic conditions between hot and cold spots. The calculation results show that the value of elevation, slope and the light intensity of the hot spots is 28.57 m, 12.67º, and 1.028, respectively. The value of cold spots is 977.06 m, 22.41º and 0, respectively. The spatial hot/cold spots analysis shows that, the cold spots are mainly distributed in the northern rocky mountain with inconvenient traffic conditions, high elevation, steep slope and less human disturbance, while the hot spots are located in the southern Guan River valley with good traffic conditions, high population density and gentle slope. It is obvious that the SD of forest aboveground biomass during the period of 1993-2013 was strongly influenced by topography and human disturbance.

Table 2 Accuracy assessment of RS estimation models by average indexes from 1993 to 2013

Models	COR	MAE	RMSE	RAE (%)	RRSE (%)
MLR	0.4835	19.2036	27.2387	86.6912	87.5655
Bagging	0.5537	17.9291	25.6909	80.9307	82.5801
k-NN	0.6596	13.2913	18.9862	59.9777	60.9947
RF	0.7184	12.6119	17.9310	56.8830	57.5532

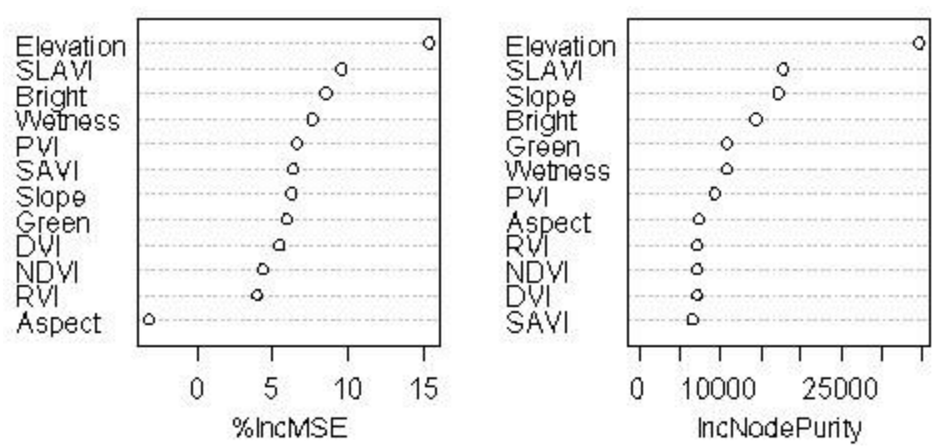

Figure 2 Mean Decrease Accuracy and Mean Decrease Gini of Environment Variables

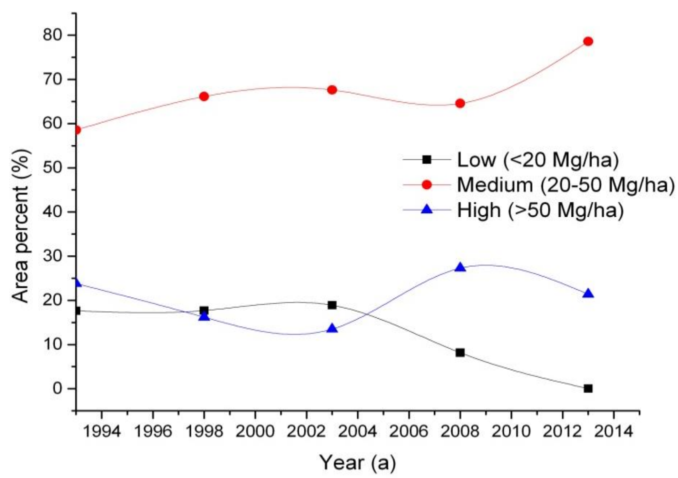

Figure 3 Trends of forest biomass in Xixia from 1993 to 2013

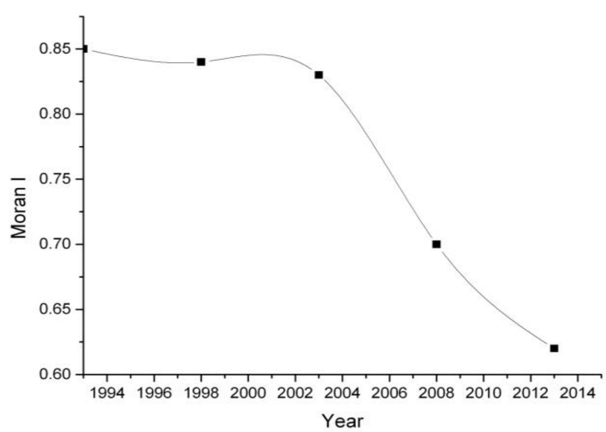

Figure 4 Dynamics of Moran *I* of forest biomass in Xixia from 1993 to 2013

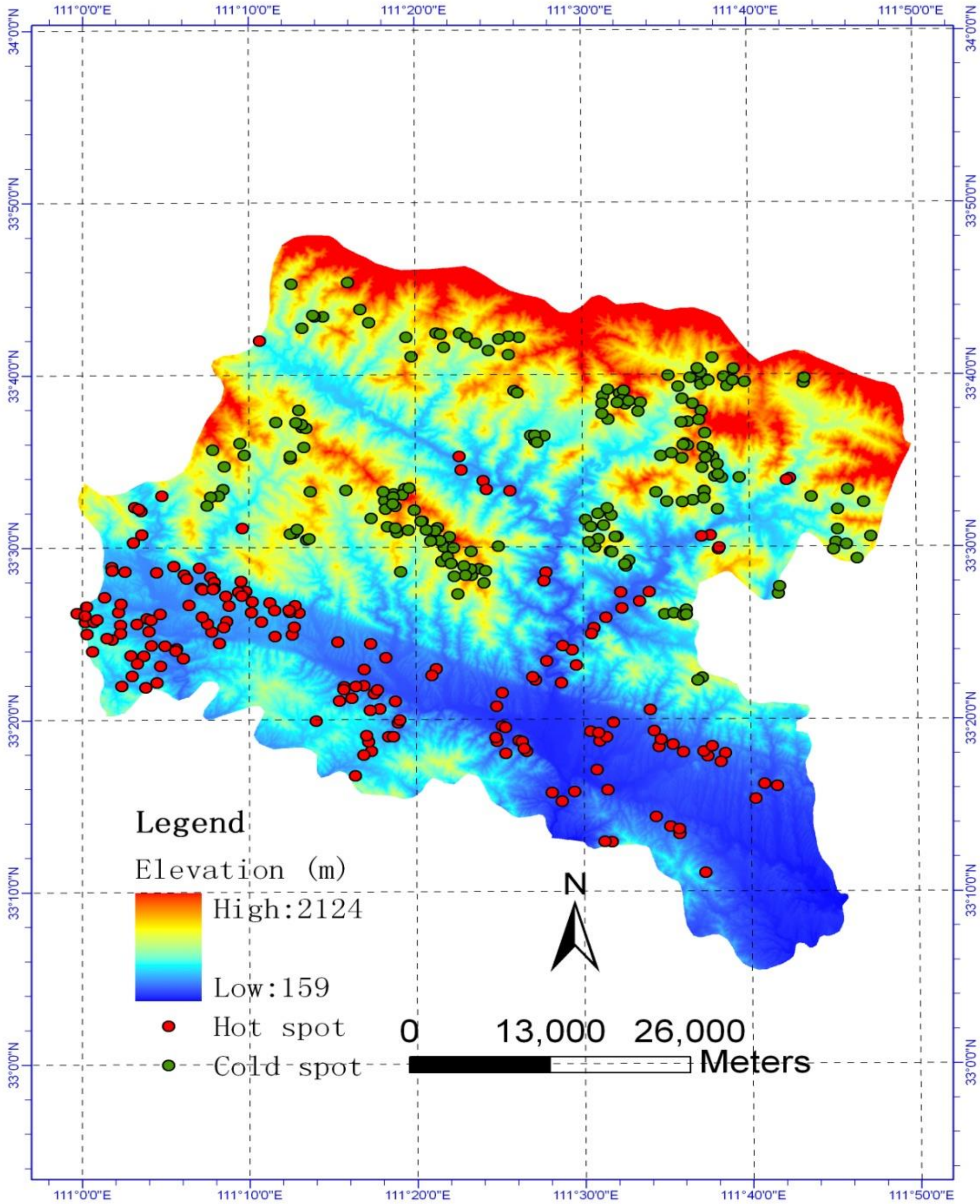

Figure 5 Cold/hot detection of SD of forest biomass in Xixia

DISCUSSION AND CONCLUSIONS

Since the establishment of continuous forest resources inventory system in 1970s, most provinces in China have carried out at least seven forest resource continuous inventories. However, the sampling population of these inventories is established at the provincial scale and forest parameters estimation results can no be allocated to counties in the province. Incomplete biomass research documents, lack of enough field survey data are the two difficulties when we monitor forest carbon sink at county level. Under this situation, modeling forest biomass by combing Landsat TM/ETM+/OLI images and fixed forest inventory plots data will be an economical flexible method with sound scientific basis.

Our studies have shown that, among the four models, the prediction accuracy of random forest is the highest, followed by k-NN method and bagging method, while the prediction accuracy of multiple linear regression is the lowest. The terrain factors including elevation and slope, soil conditions (e.g. brightness, wetness), the vegetation index (vertical vegetation index, effective leaf area index), altogether six factors are the six enforcing variables impacting regional forest biomass.

Our studies show that: during 1993 to 2013, the unit forest biomass in study area decreased in the first decade (1993-2003) and then increased in the second decade (2003-2013).The area percent of forest with high biomass (>50Mg/ha) presented a same changing trend with the unit forest biomass, while area percent of the forest with low biomass (<20Mg/ha) changed in the opposite direction. The area percent of the forest with medium biomass (20-50Mg/ha) showed a trend of steady increasing from 1993 to 2013. The forest with lower SD of aboveground biomass is mainly distributed in the northern Rocky Mountains with inconvenient traffic conditions, high forest cover and less human disturbances, while the forest with higher SD of biomass is located in the southern Guan River valley with good traffic conditions, high population density and gentle slope.

Comparative analysis shows that the forest biomass in the study area is lower than that of the average value of Henan Province in the same period, and it is much lower than the average value at the national scale. For example , the average forest aboveground biomass of Xixia County in 2008 was 36.89 Mg/ha, the average value in Henan Province was 59.9 Mg/ha, while the national average value was 85.64 Mg/ha. The main reason lies in the low forest stock volume and the poor quality of forest stand in the study area [16].In 2008 , the forest stock volume in Xixia was 31.8 m^3/ha, while the province's and national average stock volume was 42.45 m^3/ha and 78 m^3/ha, respectively. That is to say, the unit forest stock volume in study area was only equivalent to 74.91% of the provincial level, 44.77% of the national level. Research shows that forest management policies such as Mountain Distribution to Every Family, Farmland Conversion into Forest, Natural Forest Conservation, are the major drivers of the change of forest aboveground biomass in Chinese collective owned forest region. The acceleration of urbanization and the ecological corridor construction will have a profound impact on the dynamics of forest biomass in the collective forest region. With the successful implementation of forest ownership reform, the forest biomass in the study area will increase steadily, and the aggregation of forest biomass spatial distribution will be further weakened.

In the process of modeling of four remote sensing estimation models, there is a common phenomenon of over-fitting. For random forest which shows the best prediction performance, the prediction accuracy (COR) using all the modeling data as training set is 0.9685, while the average prediction accuracy (COR) using ten-fold cross validation is 0.7184. There may be four reasons for over-fitting: (1) Forest resources inventory is labor intensive field work, there maybe measurement errors during the process of data collection. (2) In the model of volume- biomass conversion, some dominant species (groups) adopted the coefficients of the equations at the national scales, which may not adapt to the actual growth of forest in the study area. (3) Among 12 variables, there may be a strong correlation between some variables, thus reducing the prediction accuracy of the models. (4) The fixed sample plot size (25.82m * 25.82m) and the spatial resolution of multi-spectral bands of remote sensing image (30m * 30m) is not consistent, which may has some impacts on the prediction accuracy.

MATERIALS AND METHODS

Data sources and preprocessing

The main data source used in this article include: (1) Fixed forest continuous inventory plot data with size of 25.82 m×25.82m in 1993, 1998, 2003, 2008 and 2013. The main technical standards of field work for each period were all the same. The survey factors included more thon 60 variables, such as geographic coordinates, site conditions, forest growth status, and so on. (2) Landsat TM/ETM+/OLI images in 1993, 1996, 2003, 2008 and 2013 with a spatial resolution of 30m×30m in multi-spectral bands and 15m×15m in panchromatic bands downloaded from website of USGS,US (http://glovis.usgs.gov/). The path/row number of the images is 125/037 and only the images acquired during growing season (May to October) with cloud cover below 20% were collected. Due to the lack of suitable images which meet requirements motioned above in 1998, a scene of Landsat TM image in 1996 instead was gathered. (3) Digital elevation model (DEM) of study area with a spatial resolution of 30m ×30m gathered from the website of Global Land Cover Facility of University of Maryland, US (http://www.landcover.org/ data/). (4) DMSP/OLS data in 1993, 1998, 2003, 2008 and 2013 from NOAA, US (http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html). Different from AVHRR, SPOT and Landsat sensors, DMSP/OLS can be used to monitor the radiation characteristics of the sun's light. The sensor can work in the night, and can detect the city lights and even the low intensity light from small scale residents, traffic flow and so on. Research shows that the light intensity is positively related to the regional population density and economic development level. Therefore, it is often used as an indicator of regional human disturbance [13].

Research Methods

Selection of Vegetation Index

In mountainous areas with big elevation gradient, forest biomass is closely related to the growth status of vegetation, terrain factors of (e.g. elevation, slope), and soil fertility and hydrological conditions [14]. Therefore, the ecological environment factors of forest aboveground biomass are extracted from these three aspects. The terrain factors include elevation, slope and aspect. The growth condition of forest vegetation is measured mainly through vegetation index. Different vegetation index measures the relationship between vegetation growth conditions and spatial distribution from different angles, and each index has its advantages and disadvantages. For example, NDVI is sensitive to the growth status and spatial distribution of green plants, but it is also strongly influenced by soil background. So, NDVI is suitable for the area with less vegetation. RVI is a sensitive indicator of the growth condition of forest, and has a high correlation with chlorophyll content. Same to the NDVI, RVI can be used to detect and estimate the plant biomass, but it is greatly influenced by vegetation coverage. When the vegetation coverage is high, RVI is very sensitive to vegetation. When the vegetation coverage is <50%, the sensitivity is significantly decreased. Because of the large size of the study area, the versified forest types and growth conditions, several vegetation indices were calculated to reflect the vegetation growth and spatial distribution of local forest. In the paper, normalized difference vegetation Index (NDVI), difference vegetation index (DVI), ratio vegetation index (RVI), vertical vegetation index (PVI), soil adjusted vegetation index (SAVI), greenness (Greenness) and effective leaf area vegetation index (SLAVI) were selected as the independent variables. At the same time, 2 variables of wetness and brightness derived from Tasseled Cap Transform of remote sensing images were used as soil factors. Together, 3 terrain factors, 7 vegetation factors, and 2 soil factors were combined together to be acted as the independent variables to predict forest aboveground biomass in Xixia County. With the processing level of L1T, all the collected remote sensing images have been geometrically corrected and orthorectified. Therefore, image preprocessing steps mainly include atmospheric correction, the subset of study area, and cloud removal. Atmospheric correction was conducted using the FLAASH atmospheric correction tool of ENVI 5.0, cloud detection and removal were completed through the added on Haze Tool of ENVI. Then, 7 vegetation indices and 2 soil indices were figured out directly using ENVI 5.0. Finally, slope and aspect of the study area were generated using the spatial analysis toolbox of ArcGIS 9.3.

Estimation of Aboveground Biomass

The forest aboveground biomass of a fixed plot can be accurately estimated by summing the biomass of all the individual trees using allometric equations based on the DBH and the height of the trees. Since there is no allometric equations available for the study area, the method presented by Fang et al. [15] was used to estimate the aboveground biomass of the plots. The regression equation suggested by Fang can be expressed as: $\mathrm{B} = aV + b$, where B is the biomass per unit area

(Mg/ha), V is the volume per unit area (m^3/ha), *a* and *b* are coefficients that vary with forest types . Among the 12 dominant tree species (groups) of Xixia County, the regression equations of three dominant tree species of *Cunninghamia lanceolata*, *Populus* spp and Paulownia spp. had been studied by local experts, so the equation coefficients of these three species were determined using research results of local experts. Compared with the research results of Fang et al. [15], the coefficients of the mixed coniferous forests, mixed coniferous and broad-leaved forest, mixed broad-leaved forest are not in conformity with the actual situation of Henan Province [16].For example, the coefficient of b in Fang's research is as high as 91.0013, but the average volume stock of mixed broad-leaved forests in the province is only 6.67 Mg/ha. Therefore, the coefficients of the three forest types were derived from the research results of Zeng [17].For the remaining dominant tree species, the coefficients are all derived from Fang.'s research (Table 1).

In forest resource continuous inventory, only the stock volume of arbor trees, open forest, scattered trees, and trees near the village, adjacent to the house, street and water, is calculated. There is no stock volume available for cash forest, shrubs and bamboo forest. In the paper, forest aboveground biomass for these forest types was estimated using non-stock conversion method suggested by Fang et al [15]. In this study, the average aboveground forest biomass of cash forest is 23.170 Mg/ha ,while the average aboveground biomass of shrub forest is 19.176 Mg/ha [15].

Modeling methods

There are many kinds of remote sensing based models to predict forest aboveground biomass. However, each model has its advantages and disadvantages. To avoid big estimation errors caused by using a single model, four remote sensing based models namely multiple linear regression (MLR), k-Nearest Neighbor k-NN), bagging (Bagging) and random forest (RF) were established in the Waikato Environment for Knowledge Analysis (Weka 3.7). Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from user's own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Multiple linear regression (MLR) is a multivariate statistical technique for examining the line correlations between two or more independent variables and a single dependent variable. Due to its good interpretability, MLR has become one of the most commonly used remote sensing based parameters inversion models. Forest biomass is closely related with plant spectral characteristics, topography, and other factors, so we can establish a MLR model based on the relationship between the forest biomass and various vegetation and soil index derived from remote sensing images [18]. As a pattern recognition method, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. It is also a type of instance-based learning, or lazy learning method. In k-NN, the nearer neighbors are always be assigned more weight and contribute more to the average than the more distant ones. The k-NN is an extremely flexible classification scheme, and does not involve any preprocessing (fitting) of the training data. This can offer both space and speed advantages in solving very large problems. Since the covariance structure of the explanatory variables can be maintained, the k-NN is widely applied to biomass mapping in forest resources survey. For example, the k-NN based research of forest biomass estimation in Newfoundland, Canada, done by Labrecque et al. [8] showed that the prediction accuracies of k-NN and multiple linear regression were similar. The research done by Ohmann and Gregory [9] showed that, in a large area with obvious biological and physical gradient variation, the auxiliary biophysical data, such as altitude and climate data, could improve the prediction accuracy of a k-NN model.

Bagging algorithm and random forest (RF) belong to the integrated learning algorithm .Bagging algorithm takes the idea of bootstrap method, using the same algorithm to handle the training samples for several times, establishing multiple independent classifiers .The final output is the vote or average value of each classifier[19]. Bagging technology as an effective multi-learner method has made many important achievements. Bagging introduces the bootstrap sampling technique into the procedure of constructing component learners, and expects to generate enough independent variance among them. Random forest is an algorithm for classification developed by Leo Breiman which uses an ensemble of classification trees. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables. Thus, random forest uses both bagging, a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned; so as to obtain low-bias trees [9].

Indicators of model validation

The correlation coefficient (COR) and root mean squared error (RMSE) of the predicted values and the measured values are often used to evaluate the prediction accuracy of the remote sensing based estimation models [20]. In addition, the mean absolute error (MAE), relative absolute error (RAE) and root relative squared error (RRSE) are also used in the evaluation of model performance by some experts [21]. In this paper, the above 5 indexes were figured out to evaluate the performance of the four models using ten-fold cross validation method. The formula of each index and its definition are as follows:

①Correlation Coefficient (COR) $\overline{y}$

$$COR = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

Where x_i is the *i*-th observed value in the field survey and y_i is the *i*-th predicted value, $\overline{x}$ and $\overline{y}$ are the average of measured values and predicted values, respectively, and *n* is the number of measurements. Correlation coefficient is an important indicator used to calculate how strong the correlation of two variables is. The dependence between two measures is obtained by dividing the covariance of the two variables by the multiplication of their standard deviations and is represented mathematically by COR(X, Y). The correlation coefficient ranges from -1 to 1. A -1 indicates perfect negative correlation, and +1 indicates perfect positive correlation.

② Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{predicted,i} - x_{observed,i})^2}{n}} \quad (2)$$

Where $x_{predicted\ i}$ is the i-th predicted value, $x_{observed\ i}$ is the i-th observed value in the field survey data, and n is the number of measurements. The root mean squared error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed in the field survey. These individual differences are also called residuals, and the RMSE severs to aggregate them into a single measure of predictive power. The lower the RMSE value is, the higher the prediction accuracy of the model.

③ Mean absolute error (%MAE)

$$MAE = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n} \quad (3)$$

Where x_i is the i-th predicted value, $\bar{x}$ is the average of the predicted value, and n is the number of the measurements. The mean absolute error measures how far estimates or forecasts differ from actual values. It is most often used in a time series, but it can be applied to any sort of statistical estimate. In fact, it could be applied to any two groups of numbers, where one set is “actual” and the other is an estimate, forecast or prediction. The lower the MAE value is, the higher the prediction accuracy of the model.

④ Relative absolute error (%RAE)

$$RAE = \sum_{i=1}^{n} \frac{\left| x_{observed,\ i} - x_{predicted,\ i} \right|}{x_{observed,\ i}} \times 100\% \quad (4)$$

Where the meanings of $x_{predicted\ i}$, $x_{observed}$ and n are same with RMSE. The relative absolute error is the ratio of the measured absolute error and the measured true value, often expressed as a percentage. In general, the relative error can reflect the credibility of the measurement. The lower the RAE is, the higher the prediction accuracy of the model.

⑤ Root relative squared error (%RRSE)

$$RRSE = \frac{1}{n} \sqrt{\sum_{i=1}^{n} \left[\frac{\left| x_{observed,\ i} - x_{predicted} \right|}{x_{predicted,\ i}} \right]^2} \times 100\% \quad (5)$$

Where the meanings of $x_{predicted\ i}$, $x_{observed}$ and n are same with *RMSE*. The root relative squared error is the ratio of the root mean squared error and the measured true value.

REFERENCE

1. YANG, Y. P., 2010. Global climate change and forest carbon sink [J]. *Journal of Sichuan Forestry Science and Technology*, 31(1), pp. 14-17. (in Chinese)
2. Luyssaert, S., Ciais, P., Piao, S.L., et al., 2010. The European carbon balance. Part 3: Forests [J]. *Global Change Biology,* 16, pp. 1429-1450.
3. ZHANG, Z., GAO, L., 2007. Analysis on historical changes of property right system of collective woodlands in southern China [J]. *Journal of Fujian Forestry Science Technology*, 34(1), pp. 170-173. (in Chinese)
4. M, Main-Knorn., Cohen, W.B., Kennedy, R.E., et al., 2013. Monitoring coniferous forest biomass change using a Landsat trajectory-based approach [J]. *Remote Sensing of Environment,* 139, pp. 277-290.
5. CAO, J.X., TIAN, Y., WANG, X.P., et al., 2009. Estimation methods of forest carbon sink and development trend [J]. *Ecology and Environmental Sciences*, 18(5), pp. 2001-2005. (in Chinese).
6. Labrecque, S., Fournier, R.A., Luther, J.E., et al., 2006. A comparison of four methods to map biomass from Landsat TM and inventory data in western Newfoundland [J]. *Forest Ecology and Management,* 226, pp. 129-144.
7. Blackard, J.A., Finco, M.V., Helmer, E.H., et al., 2008. Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment,* 112, pp. 1658-1677.
8. Moisen, G.G., & Frescino, T.S., 2008. Comparing five modeling techniques for predicting forest characteristics [J]. *Ecological Modeling*, 157, pp. 209-225.
9. Breiman, L., 2001. Random forests [J]. Machine Learning, 45, pp. 5-32.
10. XING, S.L., ZHANG, G.L., LIU, H.T., et al., 2004. The estimating model of Larix gmelinii forests biomass using Landsat ETM Data [J]. *Journal of Fujian College of Forestry,* 24(2), pp. 153-156. (in Chinese)
11. GUO, Q.X., ZHANG, F., 2003. Estimation of forest biomass based on remote sensing [J]. *Journal of Northeast Forestry University*, 31(2), pp. 13-16. (in Chinese)
12. CHEN, E.X., LI, Z.Y., WU, H.G., et al., Forest volume estimation method for small areas based on k-NN and Landsat Data [J]. *Forest Research,* 21(6), pp.745-750. (in Chinese)
13. HE, C.Y, SHI, P.J, LI, J.G., et al., 2006. Research on the process of urbanization in China in 1990s based on DMSP /OLS night light data and statistical data [J]. *Science Bulletin*, 51(7), pp. 856-861. (in Chinese)
14. XU, P., XU, T.S., 2008. Research on the estimation method of carbon storage of the forest in Gaoligongshan Nature Reserve of Yunnan Province [J]. *Forest Resources Management*, 1, pp. 69-73. (in Chinese)
15. Fang, J.Y, Chen, A.P, Peng, C.H., et al., 2001. Changes in forest biomass carbon storage in China between 1949 and 1998 [J] .*Science*, 292, pp. 2320-2322.
16. GUANG, Z.Y., 2006. Study on forest biomass and productivity in Henan [J]. *Journal of Henan Agricultural University,* 40(5), pp. 493-497.
17. ZENG, W.S., 2005. Research on forest biomass and productivity in Yunnan [J]. *Central South Forest Inventory and Planning*, 24(4), pp. 1-13.
18. WANG, H.Y., GAO, Z.H., WANG, F.Y., et al., 2010. Estimation of vegetation biomass using SPOT 5 satellite images in Fengning County, Hebei Province [J]. *Remote Sensing Technology and Application,* 25(5), pp. 639-646.
19. Breiman, L., 1996. Bagging predictors. Machine Learning, 24(2), pp. 123-140.
20. Zheng, B., Agresti, A., 2000. Summarizing the predictive power of a generalized linear model [J]. Statistics in Medicine, 19. pp. 1771-1781.
21. Deng, S.Q., Katoh, M., Guan, Q.W., et al., 2014. Estimating forest aboveground biomass by combining ALOS PALSAR and WorldView-2 Data: A Case Study at Purple Mountain National Park, Nanjing, China [J]. *Remote Sensing,* 6(9), pp. 7878-7910.
22. Griffith, D.A., 1987. Spatial autocorrelation: A primer resource publications in Geography [M]. Washington: Association of American geographers.
23. Anselin, L., 1995. Local Indicators of Spatial Association – LISA [J].

Geographical Analysis, 27(2), pp. 93-115.

Article Key words
Forest aboveground biomass, Landsat, Ground inventory.

Acknowledgments
This research was partially supported by Ural Forest State University. We thank our colleagues from Nanjing Forestry University who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations or conclusions of this paper.

Article History
Received: 05 March 2018
Accepted: 27 April 2018
Published: July-September 2018

Citation
Seyed Omid Reza Shobairi, Vladimir Andreevich Usoltsev, Viktor Petrovich Chasovskikh, Mingyang LI. Exploring forest aboveground biomass estimation using landsat, forest inventory and analysis data base. *Climate Change*, 2018, 4(15), 632-641